Improving the Race Edit in the Consumer Expenditure Survey[1] October 2013

Sharon Krieger, Barry Steinberg, and David Swanson
U.S. Bureau of Labor Statistics, Office of Prices and Living Conditions
2 Massachusetts Avenue, NE, Washington, DC 20212
Steinberg.Barry@bls.gov

Abstract

In recent years there has been concern about the declining coverage rate of the Black/African American population in the Consumer Expenditure Survey (CE). This demographic group seemed to be undercounted in the survey and the problem was growing over time. Analysis of the data found the biggest source of the undercount was among nonrespondents – people who were unable to be contacted or who refused to participate in the survey. CE has various ways of getting the race of nonrespondents, but as a last resort about 20% of them have to be randomly assigned to a category. Research showed this particular edit process was systematically placing too few nonrespondents into the Black/African American category. This paper describes an improvement CE recently made to the random assignment process. The probabilities used to assign nonrespondents to a race category were based on the racial distribution of respondents, but research showed that respondents and nonrespondents had different distributions. Simply changing the source of data used to generate the racial probabilities from the survey's respondents to the most recent Decennial Census improved the results.

Key words: Coverage Rate, Nonrespondent, Race, Imputation

1. Introduction

The Consumer Expenditure (CE) survey is a nationwide household survey conducted jointly by the U.S. Bureau of Labor Statistics and the Bureau of the Census to find out how Americans spend their money. The target population is the U.S. civilian noninstitutional population. A survey's coverage rate is the proportion of its target population that is contained in its sampling frame, and CE's coverage rate for the Black/African American population fell to around 72% in recent years. That means only 72% of Black/African American people in the U.S. civilian noninstitutional population were represented in the CE survey. Concern was raised about this low coverage rate because that group was undercounted and the problem was growing over time.

Research was conducted to find the source of the undercount and the biggest source was found to be in the nonrespondent portion of the sample. In particular, CE's edit process was undercounting Black/African American people by systematically placing too few nonrespondents with unknown or ambiguous races into the Black/African American race category. That meant Black/African American people may have been adequately represented in the CE survey in spite of their low coverage rate because too often the edit process put them in the wrong category.

---

[1]Any opinions expressed in this paper are those of the authors, and do not constitute policy of the Bureau of Labor Statistics.

As a result of this research, three actions were taken to improve the random assignment process:

- *The source of data for generating racial proportions was changed from CE's respondents to the latest Decennial Census.*
- *The level of geographic detail used to create racial proportions was refined from the four regions of the country (Northeast, Midwest, South, and West) to individual counties.*
- *Household tenure (owner, renter) was added to the process by creating racial proportions at the county/tenure level.*

This paper examines the degree to which these changes in CE's edit process improve the overall count of Black/African American people as well as the accuracy of individual imputations. While statistically significant differences were not obtained for all of these actions, there was some improvement. The changes help to correctly put more Black/African American people into their right race category, which is improving CE's household weights and expenditure estimates.

2. Background

As mentioned above, the CE survey is a nationwide household survey conducted jointly by the U.S. Bureau of Labor Statistics and the Bureau of the Census to find out how Americans spend their money. CE data have many uses. They are used as the expenditure weights for the Consumer Price Index; they are used to define poverty thresholds for the U.S. government's new Supplemental Poverty Measure; and they are used to calculate the U.S. Gross Domestic Product.

The CE actually consists of two sub-surveys, each with its own questionnaire and sample: the Quarterly Interview survey and the Diary survey. The purpose of the Quarterly Interview survey is to obtain detailed expenditure data on large items such as property, automobiles, and major appliances; and also on expenditures that occur on a regular basis, such as rent, utility bills, and insurance premiums. The purpose of the Diary survey is to obtain detailed expenditure data on small, frequently purchased items such as food and apparel. This paper focuses on the Quarterly Interview survey.

The Quarterly Interview survey is a nationwide rotating panel survey in which approximately 12,000 addresses are contacted each calendar quarter of the year. One-fourth of the addresses contacted each quarter are new to the survey. Of those 12,000 addresses, approximately 9,600 have occupied housing units, and usable interviews are obtained from approximately 7,100 of them. After an address has been in the sample for four consecutive quarters, it is dropped from the survey and a new address is selected to replace it.[2]

3. Data Description

The research presented in this paper is based on the Quarterly Interview survey data collected from 2005 through 2008. The dataset includes both respondents (households with completed interviews) and nonrespondents (households that could not be contacted

[2] The survey actually has five interviews with 15,000 addresses. The other interview is a "bounding" interview that provides baseline data which are not used to compute the survey's published expenditure estimates. The bounding interview was not used in this study and it is not reflected in the sample sizes above.

or refused to give interviews). The dataset excludes residential addresses that are not occupied and nonresidential addresses that are not meant to be in the sampling frame. The dataset contains information from 147,886 interviews and interview attempts conducted at 49,892 households with each household providing up to four interviews. The 147,886 records consist of 110,524 respondents and 37,362 nonrespondents.

4. How CE Gets the Race of Nonrespondents

It is important for a survey to know some information about its nonrespondents, and race is one of the characteristics CE requires of all households in its sample. Knowing this information is important because it is used to adjust the sampling weights of respondents to properly account for the nonrespondents. The race of respondents is almost always available because it is reported by someone in the household, but the race of nonrespondents is frequently unavailable precisely because of their nonrespondent status. For this reason CE uses some common approaches to ensure that all of its nonrespondents are assigned to a race category.

One approach is for the survey's field representatives to observe the information directly when trying to persuade the household to participate in the survey, and another approach is to ask one of their neighbors. Approximately 40% of CE's nonrespondents are assigned to a race category in these two ways. Two other ways are to use the information obtained in a partial interview, and to carry forward the information reported in a previous interview. Approximately 40% of CE's nonrespondents are assigned to a race category in these two ways. The other 20% have an unknown race category.

For the 20% of nonrespondents who still have an unknown race category, a race category is randomly assigned to them. It is assigned or "imputed" to them using probabilities generated from the racial distribution found in a specified data source. The data source used to generate these probabilities is the focus of this paper.

5. Summary of CE's Edit Process for Race

CE's edit process for race occurs in two stages. The first stage edits race at the member level, and the second stage edits race at the household level.

The goal of the first stage edit (member race edit) is to assign every household member to one of six race categories: White, Black/African American, Native American, Asian, Pacific Islander, or Multi-race. These are the official race categories used by the federal government. In this stage the following edit occurs for any household member having a missing race value:

1. IF the member is Hispanic, THEN their race is set equal to White.
2. ELSE IF the member is the reference person,[3] THEN their race is set equal to the race of another household member if available.[4]
3. ELSE their race is randomly chosen from one of the six race categories.

The goal of the second stage edit (household race edit) is to assign every household to a race category of either Black/African American or Non-Black.[5] This is done for all

[3] The reference person is the first person mentioned by the respondent when asked to "Start with the name of the person or one of the persons who owns or rents this home."

[4] CE has a hierarchy for determining which household member's race to use when the reference person's race is missing: Their spouse, child, in-law, grandchild, brother or sister, etc.

households, both respondents and nonrespondents, in order to properly adjust their sampling weights to account for nonresponse. If the race of the reference person is Black/African American, then the race of the household is Black/African American. If the race of the reference person is White, Native American, Asian, or Pacific Islander, then the race of the household is Non-Black. Otherwise, the following edit occurs:

1. IF the reference person is Multi-racial, THEN the household's race is set equal to Non-Black.
2. ELSE IF no member records are available for the household, THEN the household's race is set equal to Black/African American or Non-Black based on the individual race category reported by field representatives.
3. ELSE the household's race is randomly chosen as Black/African American or Non-Black.

6. Multi-race Assignment

The way respondents are asked about their race allows them to identify themselves by more than one race category.[6] For example, a person can say they are both White and Black/African American. When this happens they are assigned to the Multi-race category during CE's member edit process, and then later they are assigned to the Non-Black category even if Black/African American is one of the races they reported.[7]

One federal government survey (NHIS) tried to find out which race category Multi-racial people identified with most by asking them a follow-up question on which single "primary" race best represents them. By asking this follow-up question they found most Multi-racial people with Black/African American as one of their races chose Black/African American as their primary race category. This suggests that CE's procedure of putting all Multi-racial people in the Non-Black category may not be ideal. It shows one way that some Black/African American people can be misclassified by putting them into the Non-Black category. Procedures like this may be contributing to CE's problem.

7. Three Actions Implemented to Improve the Race Edit

As mentioned in the introduction, there were three actions implemented to improve the random assignment process. The first one changed the source of data for the racial proportions from CE's respondents to the latest Decennial Census. The second one changed the level of geographic detail used for the proportions from the four regions of the country to individual counties. The third one added household tenure to the process. The next three sections provide a summary of each action.

---

[5] Non-Black is used throughout this paper for any individual or household that is not Black/African American.

[6] This is how the race question is asked in the CE Interview survey: "Please choose one or more races that (name)/you considers/consider himself/herself/yourself to be./you consider (child's name) to be: *1. White, 2. Black or African American, 3.American Indian or Alaska Native, 4. Asian, 5. Native Hawaiian or other Pacific Islander, 6. Other-- Specify, 7. Don't Know. * Probe if necessary. * Enter all that apply."*

[7] People are asked to choose a race because race is now considered to be a social construction rather than an innate biological characteristic. People are free to choose whatever category they want, and they are free to choose as many categories as they want.

Action 1: Use Decennial Census Information to Randomly Assign Race

Research found that respondents and nonrespondents had different racial distributions, but the edit process used the same distribution to assign both respondents and nonrespondents with an unknown race to a race category. Specifically, the racial distribution of CE's respondents was used to randomly assign them to a race category. The problem with this was the percent of respondents who are Black/African American was less than the percent of nonrespondents who are Black/African American which caused too few nonrespondents to be assigned to the Black/African American category.

Table 1 shows the percent of Black/African Americans for CE's respondents and nonrespondents who reported their own race. Almost all respondents report their own race, with 12.2% of them reporting their race to be Black/African American. Many nonrespondents also report their race. In some cases, a partial interview was conducted in which they provided their race before prematurely ending the interview, and in other cases they reported their race during a previous interview. Black/African Americans households represent 14.3% of such nonrespondents. The table also shows the percent of Black/African Americans who reside in the 461 counties that are part of CE's sample according to the U.S. Census Bureau. The percents in this table show the distribution of CE's nonrespondents is closer to that of the Decennial Census than to CE's respondents, which is the reason the decision was made to use racial information from the Decennial Census for the probabilities.

Table 1. Percent of Black/African Americans in CE When their Race Is Known

Region	Respondents	Nonrespondents	2000 Census
Northeast	11.1%	14.7%	15.3%
Midwest	9.3	13.5	15.8
South	19.7	19.6	20.7
West	4.7	7.8	6.4
Total U.S.	12.2	14.3	14.4

Action 2: Use Racial Proportions from Individual Counties

The second action was to change the level of geographic detail used to create racial proportions from the four regions of the country (Northeast, Midwest, South, and West) to individual counties. CE used to create its racial proportions by region of the country, with four sets of proportions being created for the four regions of the country. This was fine if accuracy below the regional level was not needed. However, it was known that the concentration of the Black/African American population varied widely within each of the regions. For example, in the South the Black/African American population ranged from 3.2% of the population in West Virginia to 36.3% of the population in Mississippi. Or for individual counties, in Maryland it ranged from 2.3% of the population in Carroll County to 63.7% of the population in Prince George's County. Thus it seemed that using racial proportions from smaller geographic areas would improve the accuracy of the imputations.

Table 2 shows the percent of nonrespondents that were in the Black/African American category after using CE's old method of imputing their race using regional proportions

from CE's respondents, and also the new percents after using Decennial Census data at the county level. Using county-level Decennial Census data instead of region-level respondent data increased the percent of nonrespondents (imputed plus nonimputed) in the Black/African American category from 12.0% to 12.41%. Since the "true" percent of Black/African Americans is 14.4%, this shows that county-level Decennial Census data worked better than the region-level data from CE's respondents.

It should be noted that only about 20 percent of nonrespondents were assigned to a race category in this edit process. The other 80 percent either reported their own race or were assigned to a race category by a field representative. For this reason, the proportion of Black/African Americans after using Decennial Census data was not as close to 14.4% as might be expected.

Table 2. Percent of Nonrespondents in the Black/African American Category After Imputing the Missing Values Using...

Region	CE Respondents by Region	Decennial Census by County[8]	2000 Census
Northeast	11.9%	12.98% ± 0.32%	15.3%
Midwest	11.4	12.21 ± 0.28	15.8
South	17.4	17.20 ± 0.33	20.7
West	6.2	6.51 ± 0.17	6.4
Total U.S.	12.0	12.41 ± 0.15	14.4

Action 3: Use Decennial Census Racial Proportions at County/Tenure Level

The third action was to add household tenure to the imputation process by creating racial proportions at the county/tenure level. Tenure means whether somebody owns or rents their home. In addition to the knowledge that the concentration of the Black/African American population varied widely within each of the regions, it was also known that Black/African Americans were more likely than Non-Black people to rent their homes. According to the American Housing Survey (2009), 53.2% of Black/African Americans were renters compared to 28.6% of Non-Black people. Thus it seemed that a higher degree of accuracy could be obtained by using racial distributions at the county/tenure level than at the county level alone.

Table 3 shows the percent of nonrespondents in the Black/African American race category after imputing missing values with three different sets of probabilities. When the probabilities were generated with CE's old method using the racial distribution of CE's respondents by region of the country, 12.0% of nonrespondents (imputed plus nonimputed) were in the Black/African American

[8] This column shows a 95% confidence interval. The imputation process uses a random number generator to assign households to a race category, and because random processes produce different results every time they are run, this process was run 30 times to examine the stability of the results. From these 30 runs the mean and standard deviation of the % Black/African American statistic was computed. The confidence intervals here show the mean value plus-or-minus 2 standard deviations.

category. When the probabilities were generated using Decennial Census data at the county level 12.41% of nonrespondents were in the Black/African American category. When the probabilities were generated using Decennial Census data at the county/tenure level 12.15% of nonrespondents were in the Black/African American category.

These results show that changing the source of the probabilities from CE's respondents to the Decennial Census increased the percent of nonrespondents in the Black/African American category. The results also show that adding tenure information decreased the percent of nonrespondents in the Black/African American category. That was an unanticipated result. The reason adding tenure information decreased the percent of nonrespondents in the Black/African American category is that CE's tenure variable is often imputed too, and the imputation process may be skewed towards assigning households with an unknown tenure to the "owner" category. However, as will be discussed in the next section, adding tenure information increased the accuracy of the imputations for individual households.

Table 3. Percent of Nonrespondents in the Black/African American Category After Imputing the Missing Values Using...

Region	CE Respondents by Region	Decennial Census by County	Decennial Census by County & Tenure	2000 Census
Northeast	11.9%	12.98% ± 0.32%	12.56% ± 0.33%	15.3%
Midwest	11.4	12.21 ± 0.28	12.03 ± 0.34	15.8
South	17.4	17.20 ± 0.33	16.79 ± 0.23	20.7
West	6.2	6.51 ± 0.17	6.48 ± 0.22	6.4
Total U.S.	12.0	12.41 ± 0.15	12.15 ± 0.12	14.4

8. Percent of Imputations Determined to be Correct

In addition to looking at the overall distributions, the accuracy of the imputations for individual households was also examined. Their accuracy could be determined because some nonrespondents became respondents in subsequent interviews, and in those subsequent interviews they reported their race. Table 4 shows how successful the imputation procedures were at assigning the "right" race to these nonrespondents.

The table shows that by using CE's region-based racial proportions from its respondents, 77.4% of the imputations matched the "right" race, while by using Decennial Census proportions by county 77.82% of the imputations matched the "right" race. This means there was a slight increase in the percent of households that were properly classified using county-level detail.

The table also shows how successful the imputation process was at assigning the "right" race to nonrespondents after using racial proportions at the county/tenure level. By using Decennial Census proportions at the county/tenure level 79.41% of the imputations matched the "right" race compared to 77.82% when using proportions by county alone. Thus adding tenure information to the county-level racial proportions improved the imputation process's success rate by an additional 1.6%. This is an improvement, but it must also be pointed out that the differences in success rates are statistically indistinguishable from each other.

Table 4. Percent of Imputations that were Eventually Found to be Right Using...

Region	CE Respondents by Region (% Right)	Decennial Census by County (% Right)[9]	Decennial Census by County & Tenure (% Right)
Northeast	80.0%	78.04% ± 2.41%	79.75% ± 2.94%
Midwest	80.4	79.63% ± 1.80	81.14 ± 2.24
South	74.8	76.15% ± 1.47	77.91 ± 1.68
West	77.3	78.24 ± 1.30	79.72 ± 1.34
Total U.S.	77.4	77.82 ± 1.30	79.41 ± 1.60

9. Results Shown in CE's Production

CE's edit process was changed in 2012 and currently uses the three procedural changes described in this paper. It uses information from the most recent Decennial Census to generate a "racial proportions file" at the county and tenure level. Table 5 shows the proportion of Black/African American people for the first six months of 2012 achieved in CE's actual production environment using the new method of generating racial probabilities. There was a noticeable increase in the overall percent of Black/African American people in CE, but more importantly, more Black/African American people were correctly placed into their right race category.

Table 5. Percent of Households in the Black/African American Category Using CE's New Set of Racial Proportions (January – June 2012)

Region	Respondents	Nonrespondents	Total Sample	2000 Census
Northeast	11.0%	15.3%	12.4%	15.3%
Midwest	10.5	12.1	11.0	15.8
South	20.4	18.1	19.8	20.7
West	4.9	5.2	4.9	6.4
Total U.S.	12.8	13.1	12.9	14.4

10. Conclusion

Concern was raised in recent years over the declining coverage rate of the Black/African American population in the CE survey. That demographic group seemed to be undercounted in the survey and the problem was growing over time. As a result of the concern, research was conducted to fix the problem. The research showed a major source of the problem was the way nonrespondents were assigned to a race category. Some of the nonrespondents were randomly assigned to a race category, and the racial proportions used to assign them were skewed towards the Non-Black population.

Further research identified three ways to fix the problem: use the racial proportions observed in the latest Decennial Census instead of CE's respondents for the probabilities;

[9] This column shows a 95% confidence interval using the same process that was used for the % Black/African American calculations in Table 2.

use racial proportions at the county level instead of at the regional level; and add tenure information to the county-level proportions.

These three things were implemented in the CE survey in 2012. The results from the first six months of 2012 are very encouraging. The percent of nonrespondents classified as Black/African Americans in the sample increased from 12.0% to 13.1% and the percent of nonrespondents assigned to the right category increased as well. Assuming the Decennial Census is the "truth," then so far CE's new edit process has been successful in placing formerly misclassified Black/African American people into their appropriate category.